

Modelarea multidimensională

Prof.dr. Manole VELICANU, lect. Mihaela MUNTEAN
Catedra de Informatica Economica, A.S.E. Bucuresti

La ora actuala organizatiile au un apetit continuu pentru date. Totusi un volum mare de date nu este relevant daca nu este organizat într-o structura care sa permita transformarea datelor în informatii utile pentru organizatie.

Modelarea multidimensională este o tehnica de proiectare logica ce permite structurarea datelor. În primul rând, modelarea multidimensională nu are legatura cu reprezentarea fizica a datelor. Nu este legata de bazele de date multidimensionale (Express, Essbase, Gentium) si nu implica stocarea informatiilor în cuburi. Conceptul de modelare multidimensională nu este nou, a aparut în perioada anilor 60, cu produsele software Express si APL. Chiar foile de calcul tabelar pot fi considerate un model multidimensional .

Cuvinte cheie: modelare multidimensională, colectii de fapte, dimensiuni, ierarhii, variabile, modelul stea, modelul fulg de zapada, depozite de date, baze de date.

1 Concepte de baza

1 Modelarea multidimensională este o tehnica de proiectare logica folosita pentru vizualizarea modelelor de date sub forma unui set de variabile cheie pentru activitatea analizata. Aceste variabile sunt descrise în functie de aspectele caracteristice ale activitatii (proces/procese) analizate.

Modelarea multidimensională faciliteaza analize complexe prin agregarea si aranjarea datelor. Spre deosebire de tehnicile folosite pentru proiectarea sistemelor tranzactionale (tehnica entitate-asociere) care folosesc concepte ca entitati, relatii între entitati, descompunere functionala; modelare multidimensională foloseste concepte ca fapte, dimensiuni, ierarhii, membri si împrastiere.

Un model conceptual multidimensional poate fi implementat în:

- baze de date relationale;
- baze de date multidimensionale;
- baze de date orientate obiect.

Modelarea multidimensională este folosita adesea în proiectarea depozitelor de date si sub unele aspecte este mai simpla si mai usor de înțeles decât modelarea entitate-asociere.

Elementele de baza ale modelului multidimensional sunt :

- colectiile de fapte;
- dimensiunile;

- membrii/atributele;
- ierarhiile;
- variabilele/indicatorii;
- gradul de împrastiere a datelor în model.

Colectiile de fapte si dimensiunile pot fi reprezentate într-o baza de date relationala ca tabele. În cel mai simplu model multidimensional, schema stea, fiecare dimensiune este descrisa prin propria tabela si colectia de fapte este stocata într-o singura tabela (tabela de fapte), indexata dupa o cheie compusa din cheile primare ale dimensiunilor. Sunt multe variante ale schemei stea, dar ele folosesc aceleasi concepte de baza si anume conceptul de tabela de fapte si tabela dimensională.

*Gradul de împrastiere a datelor este un element important în modelul multidimensional influentând dimensiunea bazei de date. De exemplu indicatorul volumul vânzarilor va avea valori valide numai pentru o mica fractiune a produsului cartezian al dimensiunilor **Produs, Client si Timp**. Daca consideram ca modelul multidimensional are 120 de clienti, 20000 de produse si 1098 de zile avem 2625200000 de combinatii posibile. Vânzarile reale pot fi numai pentru 1% sau chiar 0,1% din combinatiile posibile. Uneori acest lucru este descris ca 99% sau 99,9% împrastiere.*

â Dimensiunile

Dimensiunile au următoarele caracteristici:

- Furnizează informații descriptive despre fiecare indicator (variabilă).
- Contin în general date statice. Atributele dimensionale nu se modifică la fel de des ca atributele din tabela de fapte. Totuși uneori apar modificări ale valorilor atributelor dimensionale iar modelul dimensional trebuie să permită modificarea pentru a asigura coerența și corectitudinea datelor.
- Sunt esențiale pentru analiză. Un model dimensional ce oferă un număr mare de atribute dimensionale permite analize cât mai complexe și mai variate.
- Într-un hypercub, o dimensiune este reprezentată printr-o axă.
- Într-o schemă stea sunt tabelele care se dispun radial în jurul tabelului de fapte și se mai numesc tabele dimensionale. Aceste tabele sunt denormalizate iar cheia lor primară este o cheie simplă.

Dimensiunile trebuie să conțină un număr cât mai mare de atribute (peste 100) pentru a permite analize cât mai complexe și mai variate. Cele mai bune atribute dimensionale sunt cele textuale și se folosesc ca sursă de restricții în aplicații. De exemplu dimensiunea **Produs** poate să conțină următoarele atribute dimensionale: *categoria de produse, marca, tipul de ambalaj, mărimea ambalajului, descrierea produsului* etc.

Conceptul de *dimensiune* este fundamental pentru modelarea multidimensională. Cei mai mulți proiectanți utilizează fraza “dimensiuni comune afacerilor”. De exemplu, un model multidimensional trebuie să includă timpul ca dimensiune fundamentală. Cele mai multe instrumente de analiză multidimensională au implementat facilitățile care permit conversii de calendar și operații de agregare (zi->săptămână->lună->an) automate și chiar manipularea cu ușurință a seriilor de timp (perioade promotionale, sezoane, perioade fiscale etc.)

Alte dimensiuni “comune afacerilor” sunt : *locatia, organizatia* (ierarhia organizațională), *clientul* .

O dimensiune conține mai mulți *membri*. Un membru este un nume distinct sau un identificator folosit pentru a determina poziția unui element de date (în schema stea apare sub denumirea de atribut dimensional). De exemplu, toate lunile, trimestrele și anii formează dimensiunea **Timp** și toate orașele, regiunile și țările dimensiunea **Locatie**. Un membru poate aparține la una sau mai multe ierarhii sau poate să nu fie inclus într-o ierarhie (independent). De exemplu în dimensiunea **Produs** membru *culoare* nu este inclus în nici o ierarhie.

Se pot aranja membrii unei dimensiuni în una sau mai multe *ierarhii*. Fiecare ierarhie poate avea mai multe nivele ierarhice. De exemplu în dimensiunea **Timp** putem avea două ierarhii (figura 1). Pentru a identifica poziția unui membru într-o dimensiune se folosesc conceptele de *înălțime* și *adâncime în ierarhie*. *Înălțimea* se stabilește de jos în sus. Din acest motiv nivelul (L0) al ierarhiei reprezintă nodurile frunze ale ierarhiei (înălțimea cea mai mică). În schema stea, nivelul L0 se leagă la tabela de fapte. La ora actuală sunt două curente pro sau contra folosirii celor două concepte. *Adâncimea în ierarhie* este stabilită de sus în jos. Există posibilitatea ca doi membri care au aceeași înălțime să aibă adâncimi diferite și invers (structuri arborescente neechilibrate). Structurile arborescente neechilibrate sunt implementate cu succes în bazele de date multidimensionale, bazele de date relationale fiind necorespunzătoare.

â Colectie de fapte (facts)

Colectia de fapte este un ansamblu de date corelate constând din variabile și din date de context.

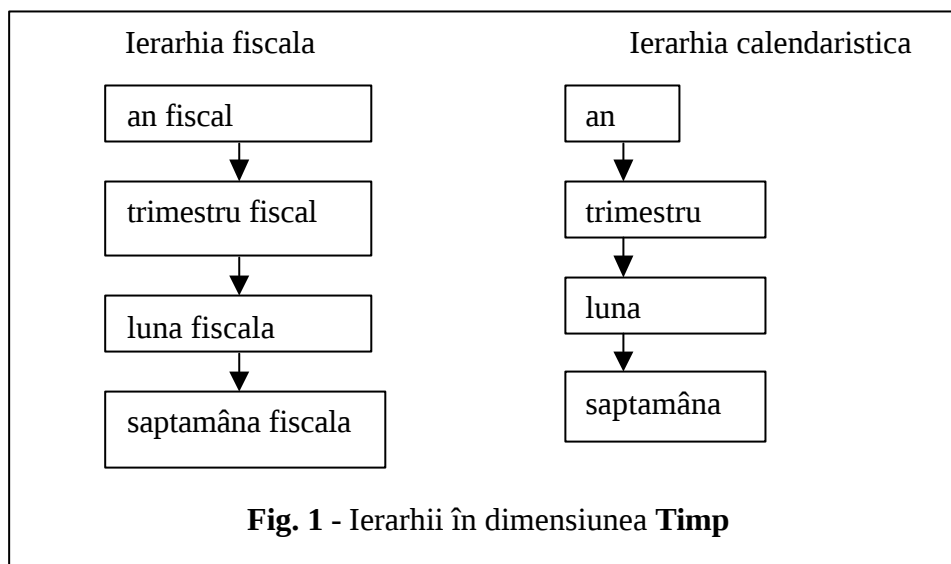
Un **fapt** poate reprezenta:

- o tranzacție a afacerii (activității) sau un eveniment (de exemplu o vânzare: ce a fost cumpărat, unde și când a avut loc vânzarea, cine a cumpărat produsul, cât s-a

platit pe unitatea de produs vândut precum și reducerile implicate în vânzare);

- starea unui obiect (de exemplu starea stocului: ce este stocat, unde este stocat și perioada de timp);

- modificările stării unui obiect dat (de exemplu modificările stocului : mișcarea produselor și cantitatea implicată în mișcare, etc)

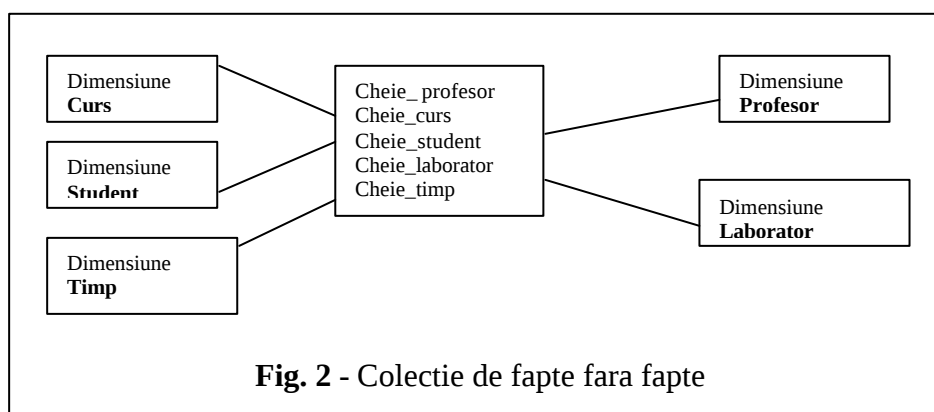


Colecțiile de fapte ce sunt asociate cu evenimente se numesc *colecții de fapte fara fapte* (factless facts), un termen folosit de Ralph Kimball deși mai bine ar fi numite *colecții de fapte fara variabile*. Evenimentele sunt adesea modelate printr-o colecție de fapte ce conține numai un număr de chei, fiecare reprezentând cheia primară a unei dimensiuni care participă la eveniment. Pentru a exemplifica se consideră cazul unei colecții de fapte în care se modelează prezenta zilnică dintr-un colegiu și care are următoarele dimensiuni:

Curs, Profesor, Student, Laborator și Timp

Colecția de fapte va conține numai cheile primare ale dimensiunilor și va permite o serie de analize interesante (figura 2):

- Care sunt cursurile cele mai frecventate/cele mai puțin frecventate ?
- Ce studenți participă la aceste cursuri?
- Care laboratoare au fost cel mai mult folosite?
- Care este rata medie de ocupare a laboratoarelor în timpul unei zile? etc.



Aceste analize presupun de fapt niste contorizari. Comanda SQL corespunzătoare

este asimetrică datorită structurii speciale a colecției de fapte. Oricare din

cele cinci chei externe poate fi folosita ca argument al functiei COUNT():

â Variabile (measures)

O variabila este un atribut numeric al unui element din colectia de fapte, un indicator de performanta prin care se poate analiza performanta/neperformanta activitatii modelate.

Valorile unui indicator se modifica continuu. Pentru fiecare combinatie posibila între dimensiuni, exista sau nu o valoare corespunzatoare a indicatorilor. Nu orice atribut numeric este un indicator. De exemplu, *dimensiunea ambalajului* este un atribut numeric si totusi nu este un indicator ci un atribut dimensional. Daca valoarea atributului numeric variaza continuu (*costul de livrare, cantitatea vânduta, volumul vânzarilor*) atunci atributul este un indicator, iar daca atributul este perceput mai mult ca o constanta (*dimensiunea ambalajului, descriere produs, culoare*) atunci este un atribut dimensional. Indicatorii pot fi clasificati în :

- indicatori de *baza* (volumul vânzarilor, cantitatea vânduta, costurile, numarul de clienti).
- indicatori *derivati* care se obtin prin combinarea indicatorilor de baza (profitul). O alta clasificare este data de Ralph Kimball [RK96] si anume dupa posibilitatea indicatorilor de a se însuma dupa dimensiuni:
- indicatori *aditivi* care se pot însuma dupa toate dimensiunile. De exemplu indicatorul *volumul vânzarilor* se poate calcula pentru o categorie de produse sau pentru o anumita regiune sau pentru a anumita perioada de timp. *Volumul vânzarilor, cantitatea vânduta si costurile* sunt aditive. Aceasta înseamna ca are sens de a aduna lei sau cantitati de-a lungul oricarei combinatii timp, produs, magazin;
- indicatori *semiaditivi* care se pot însuma numai dupa unele dimensiuni. De exemplu indicatorul *numarul de clienti* este semiaditiv deoarece se poate calcula numarul de clienti într-o anumita perioada de timp sau pentru o anumita regiune, dar

nu este aditiv de-a lungul dimensiunii **Produs**. Daca pentru fiecare produs (dintr-o categorie de produse) se cunoaste numarul de clienti si dorim sa aflam numarul de clienti pentru categoria de produse, nu se pot aduna aceste numere. Rezultatul poate fi eronat, deoarece pot exista clienti, care au cumparat mai multe tipuri de produse, din categoria respectiva;

- indicatori *nonaditivi* care nu se pot însuma dupa nici o dimensiune (profitul marginal). Variabilele nonaditive pot fi combinate cu alte variabile pentru a deveni aditive.

â Tabela de fapte

O tabela de fapte ocupa locul central într-o schema stea si corespunde unei colectii de fapte din modelul dimensional. Are urmatoarele caracteristici:

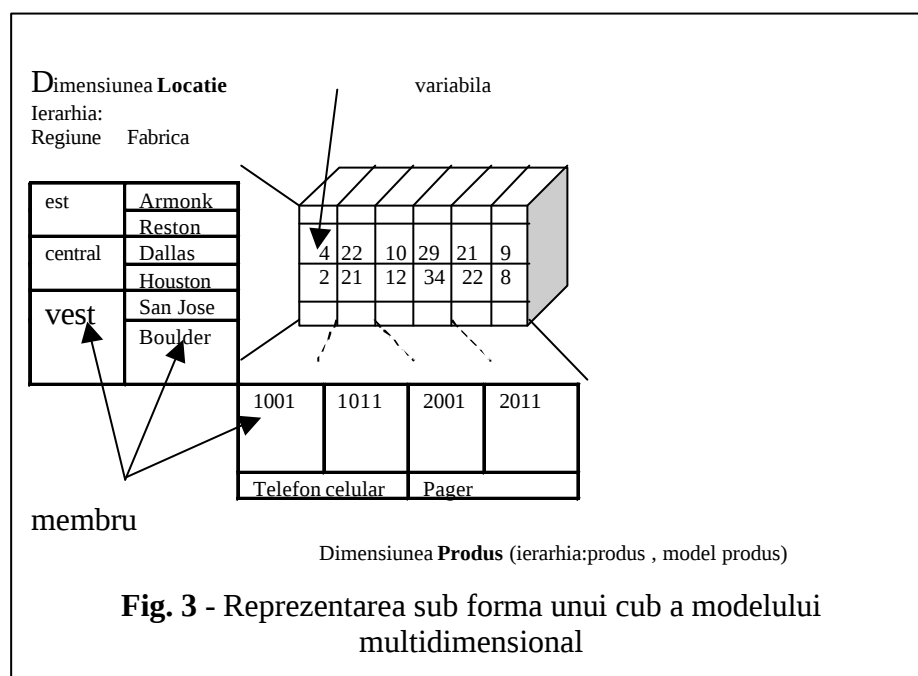
- Contine un numar foarte mare de tupluri (posibil milioane). Numarul de tupluri din tabela reprezinta de fapt produsul cartezian al dimensiunilor. De exemplu se doreste analiza activitatii de desfacere a unei firme de comert, care are un lant de 500 magazine de distributie în întreaga lume si care comercializeaza 50000 de produse în fiecare magazin. Baza de date dimensionala va stoca informatiile referitoare la tranzactiile zilnice, pe o perioada de doi ani. Tabela de fapte va contine $500 \times 50000 \times 2 \times 365$ de tupluri.
- Dimensiunea ei creste dinamic, în functie de cantitatea de date încarcate la fiecare ciclu de înprospatare a bazei de date, precum si functie de cantitatea de date istorice stocate în baza de date .
- Este tabela care reflecta performanta/neperformanta activitatii (procesului) analizate. Contine toti indicatorii de performanta importanti (în schema stea sunt de regula attribute numerice aditive).
- Cheia primara a tabelii este o cheie compusa, formata din cheile primare ale tabelilor dimensionale. Fiecare tabela de fapte dintr-o baza de date dimensionala are o cheie compusa si invers fiecare tabela care are o cheie compusa este o tabela de fapte.

- Este normalizata si realizeaza de fapt o legatura indirecta între dimensiuni.

2. Vizualizarea modelului multidimensional

Un model multidimensional se poate vizualiza cu ajutorul unui cub. Putem reprezenta un model tridimensional folosind un cub. Totuși un model multidimensional are mai mult de trei dimensiuni si este vizualizat printr-un *hypercub*. În figura 3 variabila *volumul productiei* este determinata în functie de trei dimensiuni: **Locatie**, **Produs** si **Timp**. Dimensiunea

Locatie include o ierarhie cu doua nive-luri *regiune* si *fabrica* iar dimensiunea **Produs** ierarhia cu nivelurile *produs* si *model produs*. Desi nu se reprezinta în figura, dimensiunea **Timp** se refera la anii 1999, 2000. Fiecare subcub contine *volumul productiei* pentru o anumita combinatie de valori ale dimensiunilor. De exemplu într-o anumita perioada de timp, fabrica Armonk din regiunea estica a produs 4000 de telefoane celulare, model 1001.



3. Tipuri de modele multidimensionale

În modelarea multidimensională se folosesc doua modele :

- Modelul (schema) stea..
- Modelul (schema) fulg de zapada.

Uneori *constelatia* (schema multistea) este introdusa ca o extensie a modelului stea si fulg de zapada.

â Schema stea

Este o reprezentare intuitiva a cubului de date multidimensional într-un mediu relational. Schema stea contine o tabela centrala si un set de tabele dimensionale ar-

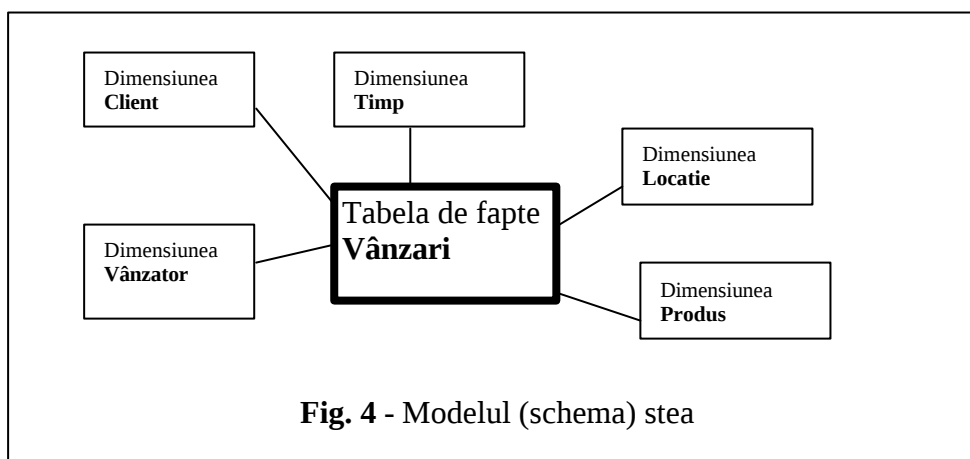
anjate într-o maniera radiala în jurul tablei centrale. Fiecare tabela dimensionala reprezinta o dimensiune a activitatii analizate, în timp ce tabela centrala contine continutul cubului de date. Schema este asimetrica. Exista o tabela dominanta în centrul schemei, singura tabela din schema cu multiple legaturi prin care se conecteaza la celelalte tabele. Aceasta tabela se numeste *tabela de fapte* (fact table). Celelalte tabele au numai o singura jonctiune prin care se leaga la tabela centrala si se numesc *tabele dimensionale* sau *dimensiuni* (dimension table). Schema stea reduce numarul de legaturi între tabele ca

urmare a procesului de denormalizare. Acestea sunt jonctiuni de egalitate. Totusi, denormalizarea determina duplicarea datelor. Figura 4 arata un exemplu de schema stea. Tabela de fapte este **Vânzari** înconjurata de dimensiunile **Timp**, **Client**, **Locatie**, **Produs**, **Vânzator**.

Schema stea are urmatoarele caracteristici:

- Tabela de fapte se leaga de dimensiuni prin joinuri de egalitate.

- Fiecare atribut din cheie primara a tabelului de fapte reprezinta cheia primara a unei dimensiuni.
- Atributele non cheie pot fi agregate. Tabelele de fapte contin numai attribute numerice
- Tabelele sunt denormalizate.

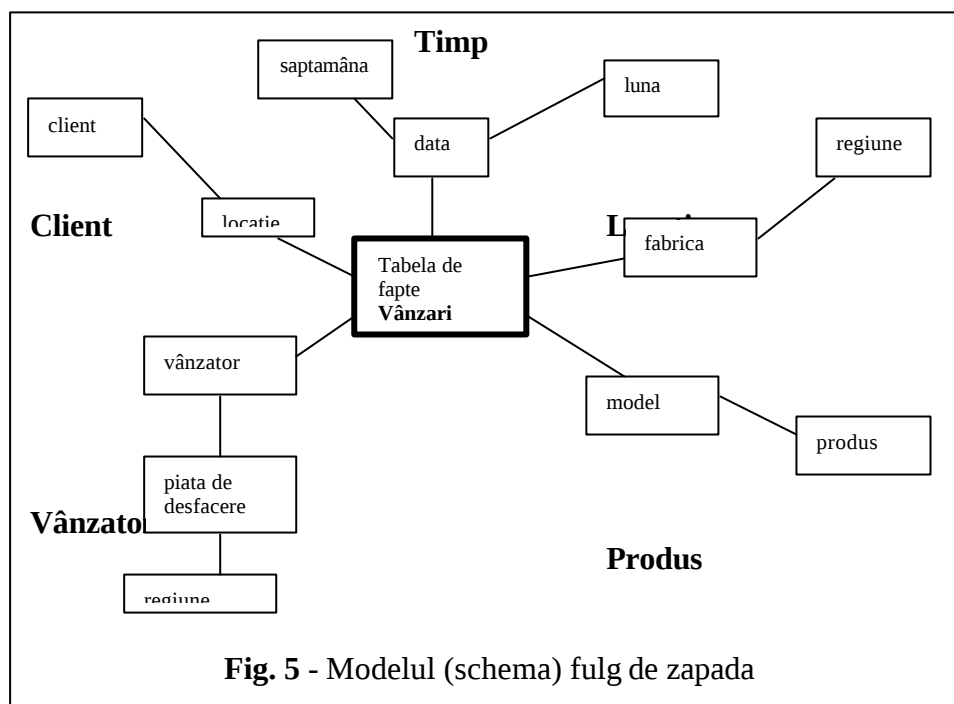


â Schema fulg de zapada

Este o varianta a schemei stea. Este rezultatul descompunerii unei dimensiuni sau a mai multor dimensiuni care au ierarhii. Relatiile de tip **(m:1)** între membrii unei dimensiuni se descompun în tabele separate formând o ierarhie de tabele. De exemplu dimensiunea **Produs** este descompusa în subdimensiunile **Model** si **Produs** (figura 5).

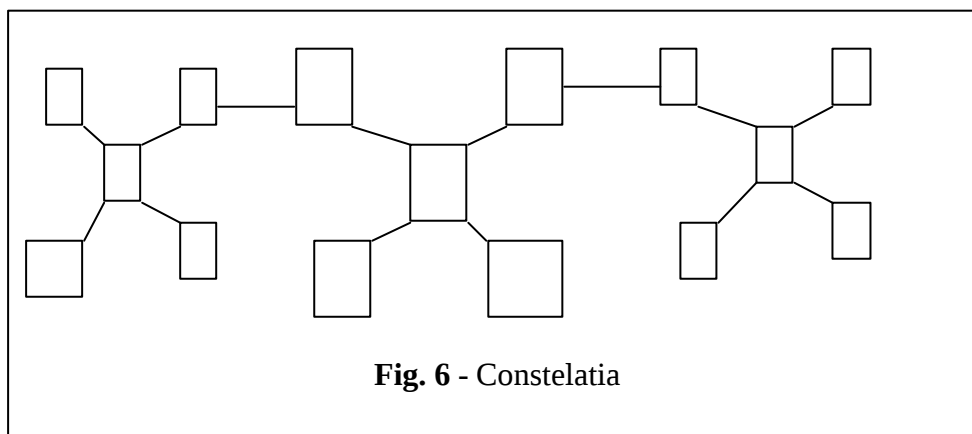
Diferenta esentiala fata de schema stea este ca dimensiunile sunt normalizate. Principala motivatie a normalizarii este spatiul

de stocare. Daca ierarhiile se pastreaza în tabele separate, se considera ca spatiul de stocare se reduce consistent. Totusi Ralph Kimball [RK96] încearca sa demonstreze contrariul. El afirma ca eforturile de a normaliza tabelele din bazele de date dimensionale, în scopul de a reduce spatiul de stocare, sunt inutile si consumatoare de timp. Normalizarea dimensiunilor reduce cu mai putin de 1% spatiul total de stocare iar vizualizarea datelor este mai dificila, implicând un numar mai mare de jonctiuni între tabele.



® O **galaxie** (constelație) este o colecție de stele și constă dintr-o stea centrală înconjurată de alte stele. Steaua centrală conține datele la nivel atomic iar celelalte

stele conțin date agregate. Steaua centrală se leagă de celelalte stele prin atribute dimensionale (figura 6).



4. Analiza comparată între modelul multidimensional și modelul entitate-asociere.

Modelul multidimensional (MMD) este proiectat pentru a suporta nevoile de raportare și cele analitice ale managerilor și poate fi descris în comparație cu modelul entitate-asociere după cum urmează (tabelul 1):

- MMD vizualizează informația din perspectiva unei “felii de timp” și nu a

tranzacțiilor atomice. Sistemele tranzacționale înregistrează evenimente sau tranzacții (de exemplu comenzile de cumpărare, facturile etc). Modelul de date multidimensional nu se concentrează pe evenimentele actuale ci pe rezultatul cantitativ al lor, într-un interval de timp cum ar fi zile, săptămâni sau ani.

- MMD este consistent la nivel global, modelul entitate-asociere este consistent numai la nivel local. Un sistem tran-

zational proiectat corect asigura consistenta datelor locale daca înregistreaza corect toate tranzactiile relevante. Daca se combina informatiile din mai multe sisteme tranzactionale, consistenta globala nu este realizata automat. Modelul multidimensional pleaca de la o consistenta globala la nivelul organizatiei.

- Sistemele tranzactionale furnizeaza informatii necesare pentru managementul operational, MMD pentru cel strategic si tactic. Modelul multidimensional este proiectat pentru a raspunde la întrebări de tipul :

“Voi avea profit în aceasta afacere sau nu?”

“Cine sunt cei mai buni clienti ai mei si de ce?”

“Ce oportunitati am pierdut?”

- Relatiile sunt modelate explicit în modelul entitate-asociere si implicit în modelul multidimensional. De exemplu relatiile explicite între clienti si comenzile de vânzare sunt distruse în proiectarea bazei de date folosind modelul entitate-asociere. În modelul multidimensional, aceste relatii sunt implicate prin existenta *faptelor* la zona de intersectie a dimensiunilor. De exemplu, daca sunt vânzari la clientul 101 pentru produsul 2345, atunci relatia între client si produs este implicita.

- Desi arata diferit, au totusi unele *asemanari*. Modelul multidimensional foloseste aceleasi concepte (entitate, relatie, atribut, cheie primara). O tabela de fapte este de fapt o entitate în care cheia primara este o combinatie de chei externe ce refera dimensiunile. De aceea putem spune ca modelul dimensional este o forma speciala a modelului entitate-asociere. Modelul entitate-asociere poate fi folosit ca punct de pornire în proiectarea modelului multidimensional.

Modelarea entitate-asociere este o tehnica de proiectare logica folosita în analiza structurala si care cauta sa elimine redundanta datelor. Astfel tranzactiile devin foarte simple si deterministe. Se pot face o serie de observatii daca privim un model entitate-asociere:

- modelul este foarte simetric. Toate tabelele arata la fel. Nu se poate stabili care tabela este mai importanta sau care este mai mare, care tabela contine indicatori de performanta ai activitatii si care nu;
- modelele complexe (pentru o întreprindere) sunt greu de urmarit si greu de înțeles de catre utilizatori;
- eficientizarea procesului tranzactional a dus la crearea unor baze de date ce nu pot fi interogate sau sunt foarte greu de interogat. Modelul entitate-asociere pentru întreprinderi are sute de entitati logice si de legaturi între acestea. Fiecare din aceste entitati, în general, apar ca o tabela, la proiectarea bazei de date. Nu se poate interoga cu usurinta un astfel de model;
- nu modeleaza real activitatea ci mai exact relatiile micro între elementele componentel;
- este foarte variabil ca structura.

Modelul multidimensional este un model **top-down** iar **modelul entitate-asociere** un model **bottom-up**. Multi proiectanti afirma ca “este mai putina informatie în modelul multidimensional” sau ca “schema stea este folosita numai pentru agregari la nivel înalt”. Ambele afirmatii sunt false. Cheia înțelegerii relatiei între modelul multidimensional si modelul entitate-asociere este ca o singura diagrama entitate-asociere (la nivel de întreprindere) corespunde la un set de diagrame multidimensionale.

Modelul multidimensional rezultat pentru o întreprindere (de dimensiuni mari) va consta din 10-25 de scheme stea similare. Fiecare schema stea va avea 4-12 tabele de dimensiuni.

Modelul multidimensional are un numar de *avantaje* importante fata de modelul entitate-asociere:

- schema stea rezista la modificarile neasteptate al cerintelor utilizatorilor. Toate dimensiunile pot fi gândite ca puncte de intrare egal simetrice în tabela de fapte. Proiectarea logica este independenta de modelele de cereri. Strategiile de interogare sunt simetrice si codul SQL-ul

generat pentru modelul multidimensional este simetric;

- schema stea reduce numarul de legaturi si permite o intretinere usoara;
- este extensibil. Se pot adauga noi tabele de fapte, noi dimensiuni, noi attribute în cadrul unei dimensiuni. Modelul multidimensional se poate extinde fara a afecta aplicatia;
- exista un numar mare de instrumente ce gestioneaza si folosesc agregatele.

Modelarea multidimensionala este aplicata în multe domenii de activitate economica (operatiuni bancare, asigurari de sanatate, comerțul cu amanuntul, activitatea companiilor telefonice, expedierea produselor etc.). Este cea mai buna tehnica de proiectare a unui depozit de date/centru de date.

BIBLIOGRAFIE

[RK96] Ralph Kimball "The Data Warehouse Toolkit. Practical Techniques for Building Dimensional Data Ware-

houses", John Wiley & Sons Inc., New York, 1996

[OC99] Oracle Corporation "The Essentials of Data Warehouse Database Design", Student Guide, 1999

[BH98] Chuck Ballard, Dirk Herreman, Don Schau "Data Modeling Techniques for Data warehousing", 1998, <http://www.redbooks.ibm.com>

[PG96] Paul Gray "The new DSS: Data Warehouses, OLAP, MDD and KDD", AIS Americas Conference, 1996

[RK97] Ralph Kimball "A dimensional modeling manifesto", DBMS Magazine, 1997

[CC97] Charles Contempo, "Coping with complex queries in a Data Warehouse", InfoDB, 1997

[LB96] Lambert Bob "Data Warehousing Fundamentals: What you need to know to succeed", Data Management Review, 1996

[RK98] Ralph Kimball "Meta Meta Data, Data", DBMS Magazine, 1998